A QUANTILE DOMAIN PERSPECTIVE ON THE
RELATIONSHIPS BETWEEN OPTIMAL GROUPING,
SPACING AND STRATIFICATION PROBLEMS

by

R. L. Eubank

Technical Report No. 163
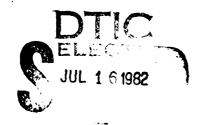Department of Statistics ONR Contract

June 1982

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

DTIC
ELEC
JUL 1 6 1982

A QUANTILE DOMAIN PERSPECTIVE ON THE RELATIONSHIPS BETWEEN OPTIMAL GROUPING, SPACING AND STRATIFICATION PROBLEMS

by

R. L. Eubank[1]

Department of Statistics, Southern Methodist University, Dallas, Texas

Running head: Relationships Between Grouping, Spacing and Stratification Problems

Abstract. The relationships between two distributions having the same solutions for problems of optimal spacing selection for the asymptotically best linear unbiased estimator of a location or scale parameter or for problems of optimal stratification for estimation of a population mean are investigated. Easily checked necessary and sufficient conditions under which two distributions have identical solutions to these problems are given in terms of their quantile and density-quantile functions. As an application of these results a quantile domain analog of a theorem due to Adatia and Chan (1981, Scand. Actuar. J., 193-202) on the equivalence of optimal grouping, spacing and stratification problems is obtained.

p 1

1. <u>Introduction</u>. For a particular absolutely continuous distribution F, assumed to depend on a location or scale parameter θ, the insightful work of Adatia and Chan (1981) has focused on the equivalence of problems of i) optimal grouping for maximum likelihood estimation of θ, ii) optimal quantile (optimal spacing) selection for the asymptotically best linear unbiased estimator (ABLUE) of θ and iii) optimal stratification for estimation of a population mean. In this paper the more general question of when <u>different distributions</u> have equivalent solutions for these problems is considered from a quantile domain perspective. We focus, initially, on the latter two problems which can be stated as follows:

<u>Problem 1</u>. Select percentile points $0 = u_0 < u_1 < \cdots < u_{k+1} = 1$, (frequently called a <u>spacing</u>) corresponding to sample quantiles which maximize the asymptotic relative Fisher efficiency of the ABLUE for θ (cf. Sarhan and Greenberg (1962, Chap. 5)).

<u>Problem 2</u>. Given strata boundaries $a = x_0 < x_1 < \cdots < x_{k+1} = b$ (where a and b are possibly infinite values which bound the support of F), a stratified random sample of size n is to be selected using proportional allocation, i.e., the "number" of sample elements taken from $(x_{i-1}, x_i]$ is $n[F(x_i) - F(x_{i-1})]$. If θ is the mean for F, the usual estimator of θ is $\hat{\theta} = \sum_{i=1}^{k+1} [F(x_i) - F(x_{i-1})] \bar{X}_i$ where $\bar{X}_i$ is the sample mean from the ith stratum. The problem is to select the boundaries to minimize the variance of $\hat{\theta}$ (cf. Dalenius (1950)). Observe that when F is a normal distribution θ is a location parameter whereas for the exponential distribution θ corresponds to a scale parameter.

Both Problems 1 and 2 are nonlinear in nature so that, as a rule, their solutions must be tabulated numerically. However, in some instances it has been possible to exploit relationships between different types of distri-

butions to obtain, for example, optimal spacings for one distribution

in terms of those for another which have already been computed

(cf. Kulldorff (1973)). When applicable, this approach can save con-

siderable time, effort and expense. The question that arises, of course,

is when and under what conditions can such tactics be expected to work or

fail. Thus, we would like easily checked conditions regarding the equi-

valence (or non-equivalence) of optimal spacing or stratification problems

for different distribution types. Motivated by such considerations we

will study the relationship between two distributions having the same

solutions for either of Problems 1 or 2. Our major results in this regard

(Theorems 1-3) are stated and discussed in the next section. It will be

seen that necessary and sufficient conditions for the solutions of any of

these problems to coincide for two distributions can be succinctly sum-

marized in terms of relationships between their quantile and density-quantile

functions. Proofs are given in Section 3 with Section 4 devoted to the

application of results in Section 2 to the optimal grouping problem.

2. <u>Optimal Spacing and Stratification</u>. Let $F_1(x;\theta_1)$ and $F_2(x;\theta_2)$

be two strictly monotone, continuously differentiable, distribution functions

(d.f.'s) which depend on parameters θ_1 and θ_2 of either the location or

scale variety. The standardized forms of these d.f.'s, corresponding

to $\theta_i = 0$ or $1(i=1,2)$ contingent on whether θ_i is a location or scale para-

meter, will be denoted as H_1 and H_2, respectively, with their associated

continuous densities written as h_1 and h_2. Thus, for example, $F_1(x;\theta)$ can be

expressed as $H_1(x-\theta_1)$, if θ_1 is a location parameter, or $H_1(x/\theta_1)$, if θ_1

is a scale parameter. Also, define the standardized <u>quantile functions</u>

(q.f.'s) and <u>density-quantile functions</u> (d.q.f.'s)

$$Q_i(u) = H_i^{-1}(u) = \inf\{x : H_i(x) \geq u\} , \quad 0 < u < 1, \quad i=1,2, \qquad (1)$$

and

$$d_i(u) = h_i(Q(u)) , \quad 0 \leq u \leq 1 , \quad i=1,2 . \qquad (2)$$

Our principal result regarding two distributions having the same optimal spacings is provided by the following theorem.

Theorem 1. Let g_i, $i=1,2$, denote either d_i or the product of d_i and Q_i, $d_i \cdot Q_i$, depending on whether θ_i is a location or scale parameter. Assume that g_i is either concave or convex, vanishes at 0 and 1 and is twice continuously differentiable on $(0,1)$ with $(g_i')^2$ and $|g_i''|^{2/3}$ integrable. Under these assumptions F_1 and F_2 will have the same optimal spacings for the estimation of θ_1 and θ_2 for all k if and only if there exists constants α, β ($\beta \neq 0$) such that

$$g_1'(u) = \alpha + \beta g_2'(u) , \quad u \in (0,1). \qquad (3)$$

To exemplify the use of Theorem 1 consider the Weibull distribution

$$F_1(x;\theta_1) = 1-\exp\{-(x/\theta_1)^\nu\} , \quad x,\nu > 0,$$

for which $H_1(x) = 1-\exp\{-x^\nu\}$, $Q_1(u) = \{-\ln(1-u)\}^{1/\nu}$ and $d_1(u) = \nu(1-u)[-\ln(1-u)]^{1-1/\nu}$. Since θ_1 is a scale parameter we use $g_1(u) = d_1(u)Q_1(u) = -\nu(1-u)\ln(1-u)$. A special case of the Weibull is the exponential distribution which corresponds to $\nu = 1$. The optimal spacings for the exponential have been tabulated and may be found, for instance, in Sarhan, Greenberg and Ogawa (1963). Taking $g(u) = -(1-u)\ln(1-u)$ it follows from (3) that these spacings are also optimal for the Weibull when $\nu \neq 1$. This relationship has also been noted by Kulldorff (1973). Other results obtained by Kulldorff (1973) also follow similarly from Theorem 1.

As another example consider the logistic distribution with

$$F_1(x;\theta_1) = [1 + \exp\{-\pi(x-\theta_1)/\sqrt{3}\}]^{-1} \ , \ -\infty < x < \infty \ ,$$

so that $H_1(x) = [1 + \exp\{-\pi x/\sqrt{3}\}]^{-1}$, $Q_1(u) = \frac{\sqrt{3}}{\pi} \ln(u/1-u)$ and

$g_1(u) = d_1(u) = \pi/\sqrt{3} \ u(1-u)$. By choosing $g_2(u) = \phi\Phi^{-1}(u)$, where Φ and ϕ

are the standard normal distribution and density function, respectively,

it is seen that the optimal spacings for location parameter estimation

for the logistic and normal distributions cannot be identical for all k.

If instead we choose $F_2(x;\theta_2) = 1-(1+x/\theta_2)^{-\nu}$, $x, \nu > 0$, it follows that

the optimal spacings for location parameter estimation for the logistic

are the same as those for scale parameter estimation in the Pareto distri-

bution if and only if $\nu = 1$.

Two distributions will be said to have the same optimal solutions

for Problem 2 if, for any set of optimal strata boundaries $\{x_{i1}\}_{i=0}^{k+1}$ for F_1,

there is a corresponding set $\{x_{i2}\}_{i=0}^{k+1}$ of optimal boundaries for F_2 which

satisfies

$$F_1(x_{i1};\theta_1) = F_2(x_{i2};\theta_2) \ , \ i=0,\ldots,k+1. \tag{4}$$

Such a definition is natural since it considers equivalence in terms of

percentage points that are not influenced by departures in the values

of x_{i1} and x_{i2} due merely to factors of location or scale. The next

theorem has the consequence that a distribution is essentially determined

by its optimal strata boundaries.

Theorem 2. If, for $i=1,2$, Q_i is square integrable and $Q_i' = 1/d_i$ is monotone

and continuous on $(0,1)$ with $d_i^{-2/3}$ integrable, then F_1 and F_2 have the same

solution for Problem 2 (in the sense of (4)) for all k if and only if there

exists constants α, β ($\beta \neq 0$) such that

$$Q_1(u) = \alpha + \beta \, Q_2(u) \ , \ u \, \epsilon \, (0,1) \ . \tag{5}$$

Theorem 1 has the implication that distributions with the same optimal strata boundaries must be members of the same family which differ by at most factors of location and/or scale. While the direct implication of this condition appears obvious the converse, although intuitive, is somewhat less transparent.

It is also reasonable to ask under what conditions the optimal spacings for one distribution can be obtained in terms of the optimal strata boundaries for another in the sense that, if $\{x_{i2}\}_{i=0}^{k+1}$ is a set of optimal boundaries for F_2, an optimal spacing for F_1 is provided by

$$u_{i1} = F_2(x_{i2};\theta_2) \ , \quad i=0,\ldots,k+1 \ . \tag{6}$$

Such conditions are provided by the following theorem.

Theorem 3. Let g_1 denote either d_1 or $d_1 \cdot Q_1$, depending on whether θ_1 is a location or scale parameter, and assume that g_1 and Q_2 satisfy the hypotheses of Theorem 1 and 2 respectively. Then Problem 1 for F_1 is equivalent to Problem 2 for F_2 (in the sense of (6)) for all k if and only if there exists constants α,β ($\beta{\neq}0$) such that

$$g_1'(u) = \alpha + \beta \, Q_2(u) \ , \quad u \, \epsilon \, (0,1). \tag{7}$$

As an illustration, note that it follows from (7) that optimal spacing selection for location parameter estimation for the logistic is equivalent to the problem of optimal stratification for the estimation of the mean of a uniform distribution on any finite interval [a,b].

3. _Proofs_. In this section Theorems 1-3 will be proven. The proofs are accomplished by a series of three lemmas.

Lemma 1. Let d _and_ Q _denote the standardized d.q.f. and q.f. for a distribution which depends on either a location or scale parameter_, θ. _Define_ g _as_ d, _for_ θ _a location parameter, or_ d·Q, _for_ θ _a scale parameter, and assume that_ g _vanishes at_ 0 _and_ 1 _and is absolutely continuous with square integrable derivative_ g'. _Then, the problem of optimal spacing selection for the ABLUE of_ θ _is equivalent to the selection of a best set of breakpoints for_ $L_2[0,1]$ _approximation of_ g' _by piecewise constants_.

Proof. See Eubank, Smith and Smith (1981).

Lemma 2. Let Q _denote the standardized q.f. for an absolutely continuous d.f._ F(x;θ) _where_ θ _is either a location or scale parameter. Assume that_ F _has a finite second moment and mean proportional to_ θ _and that_ Q _is continuous and strictly monotone on_ (0,1). _Then, Problem_ 2 _for_ F _is equivalent to selecting optimal breakpoints for_ $L_2[0,1]$ _approximation of_ Q _by piecewise constants_.

Proof. First observe that, since F has a finite second moment, $Q \in L_2[0,1]$. For the case of θ a location parameter, F may be expressed as H(x-θ), where H is the distribution function corresponding to Q. Now, for any set of strata boundaries $a = x_0 < x_1 < \cdots < x_{k+1} = b$ it has been shown by Dalenius (1950) that the variance of $\hat{\theta}$ is

$$V(\hat{\theta}) = n^{-1} \sum_{i=1}^{k+1} [H(x_i - \theta) - H(x_{i-1} - \theta)] \sigma_i^2 \qquad (8)$$

where

$$[H(x_i-\theta)-H(x_{i-1}-\theta)]\sigma_i^2 = \int_{x_{i-1}}^{x_i} x^2 dH(x-\theta) - [H(x_i-\theta)-H(x_{i-1}-\theta)]^{-1}\left[\int_{x_{i-1}}^{x_i} x dH(x-\theta)\right]^2 . \quad (9)$$

Making the change of variable $x-\theta = Q(u)$ in (9), letting $u_i = H(x_i-\theta)$ and simplifying gives

$$(u_i-u_{i-1})\sigma_i^2 = \int_{u_{i-1}}^{u_i} Q(u)^2 du - (u_i-u_{i-1})^{-1}\left[\int_0^1 Q(u) I_{(u_{i-1},u_i]}(u)du\right]^2$$

where $I_{(u_{i-1},u_i]}$ is the indicator function for $(u_{i-1},u_i]$. Thus, $nV(\hat{\theta})$ is now recognized as the squared $L_2[0,1]$ error for the approximation of Q by piecewise constants with breakpoints at $0 = u_0 < u_1 < \cdots < u_{k+1} = 1$. Observing that the u_i and x_i are uniquely defined by $\theta + Q(u_i) = x_i$ it follows that minimization of (8) with respect to the u_i's or the x_i's are equivalent problems. The case of scale parameter estimation is proven similarly.

As a result of Lemmas 1 and 2 questions concerning the equivalence of Problems 1 and 2 can now be viewed as questions regarding the equivalence of breakpoint selection problems for $L_2[0,1]$ approximation by piecewise constants. This subject is treated by the next lemma.

Lemma 3. Let m_1 and m_2 be square integrable functions and assume that, for $i=1,2$, m_i' is continuous, monotone and of one sign on $(0,1)$ with $|m_i'|^{2/3}$ integrable. Then m_1 and m_2 will have the same optimal breakpoints for $L_2[0,1]$ approximation by piecewise constants for all k if and only if there exists constants α,β $(\beta\neq0)$ such that

$$m_1(u) = \alpha + \beta m_2(u) , \quad u \in (0,1). \quad (10)$$

Proof. The sufficiency of (10) follows immediately upon noting that, in this event, the $L_2[0,1]$ errors for approximation of m_1 and m_2 are proportional with proportionality factor $|\beta|$. To establish its necessity let $\{U_k\}$ be a sequence of sets of optimal breakpoints for m_1 where $U_k = \{u_0^k, \ldots, u_{k+1}^k\}$ and $0 = u_0^k < \cdots < u_{k+1}^k = 1$. By hypothesis each U_k is also optimal for m_2. Now, as in Barrow and Smith (1978), define piecewise linear functions s_k with $s_k(u_i^k) = i/k+1$, $i=0,\ldots,k+1$. Then, using Theorem 1.1 of Burchard and Hale (1975) in conjunction with the proof of Theorem 3 in Barrow and Smith (1978), it is seen that

$$\lim_{k\to\infty} s_k(\tau) = \int_0^\tau |m_1'(u)|^{2/3} du / \int_0^1 |m_1'(t)|^{2/3} dt. \tag{11}$$

However, as the U_k are also optimal for m_2, it must be that

$$\lim_{k\to\infty} s_k(\tau) = \int_0^\tau |m_2'(u)|^{2/3} du / \int_0^1 |m_2'(t)|^{2/3} dt. \tag{12}$$

The lemma now follows by equating (11) and (12) and differentiating.

To prove Theorem 1 take $m_1 = g_1'$, $m_2 = g_2'$ in Lemma 3 and observe that the convexity or concavity of g_i is equivalent to g_i'' being of one sign on $(0,1)$. Theorems 2 and 3 can be obtained similarly by taking $m_1 = Q_1$, $m_2 = Q_2$ and $m_1 = g_1'$, $m_2 = Q_2$.

Remark. Lemmas 1 and 2 have the consequence that problems of optimal spacing and stratification, when viewed in the quantile domain, have simple geometric interpretations as piecewise constant approximation problems. Relationships between the solutions to these problems for different distributions can, therefore, often be detected by merely graphing the appropriate functions.

4. <u>Extension to Optimal Grouping</u>. The results of Section 2 can be extended to include the following optimal grouping problem considered by Kulldorff (1961).

<u>Problem 3</u>. Choose group boundaries, $a = x_0 < x_1 < \cdots < x_{k+1} = b$, that minimize the asymptotic variance of the maximum likelihood estimator of a location or scale parameter, θ, obtained using only the group boundaries and the proportion of sample elements within each group.

Theorems 1 and 3 can also be shown to apply to Problem 3 through the use of Theorem 4 of Adatia and Chan (1981). Their result states that for a given distribution Problems 1 and 3 are always equivalent (in the sense of (6)), provided the distribution satisfies certain regularity conditions specified by Kulldorff (1961, pg. 20). Using this fact and restricting attention to the special case of $F_1 = F_2 = F$, $\theta_1 = \theta_2 = \theta$, $d_1 = d_2 = d$ and $Q_1 = Q_2 = Q$, a quantile domain version of Theorem 5 of Adatia and Chan (1981) follows from Theorem 3.

<u>Corollary</u>. <u>Let</u> g <u>represent either</u> d <u>or</u> d·Q, <u>depending on whether</u> θ <u>is a location or scale parameter, and assume that</u> g <u>and</u> Q <u>satisfy the hypotheses of Theorem</u> 1 <u>and</u> 2 <u>respectively. Then Problems</u> 1-3 <u>are equivalent for all</u> k <u>if and only if there exists constants</u> α, β ($\beta \neq 0$) <u>such that</u>

$$g'(u) = \alpha + \beta \, Q(u) \quad , \quad u \in (0,1). \tag{13}$$

In comparing the Corollary with the results of Adatia and Chan observe that our approach dispenses with conditions requiring the existence of a sequence of strata boundaries, $\{B_k\}$, for which the corresponding estimators satisfy $\lim_{k \to \infty} V(\hat{\theta}_k) = 0$. We note that it follows immediately from (13) that Problems 1-3 are equivalent for either a normal or gamma distribution.

REFERENCES

Adatia, A. and Chan, L. K. (1981). Relations between stratified, grouped
and selected order statistics samples. Scand. Actuar. J., 193-202.

Barrow, D. L. and Smith, P. W. (1978). Asymptotic properties of best
$L_2[0,1]$ approximation by splines with variable knots. Quart. Appl.
Math. 36, 293-304.

Burchard, H. G. and Hale, D. F. (1975). Piecewise polynomial approximation
on optimal meshes. J. Approx. Theory 4, 128-147.

Dalenius, T. (1950). The problem of optimal stratification. Skand.
Aktuarietidskr. 33, 203-213.

Eubank, R. L., Smith, P. L. and Smith, P. W. (1981). Uniqueness and
eventual uniqueness of optimal designs in some time series models.
Ann. Statist. 9, 486-493.

Kulldorff, G. (1961). Contributions to the Theory of Estimation from Grouped
and Partially Grouped Samples, Almqvist and Wiksell: Stockholm.

Kulldorff, G. (1973). A note on the optimum spacing of sample quantiles
from the six extreme value distributions. Ann. Statist. 1, 562-567.

Sarhan, A. E. and Greenberg, B. G., eds. (1962). Contributions to Order
Statistics, John Wiley: New York.

Sarhan, A. E., Greenberg, B. G. and Ogawa, J. (1963). Simplified estimates
for the exponential distribution. Ann. Math. Statist. 34, 102-116.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 163	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A QUANTILE DOMAIN PERSPECTIVE ON THE RELATIONSHIPS BETWEEN OPTIMAL GROUPING, SPACING AND STRATIFICATION PROBLEMS		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER 163
7. AUTHOR(s) R. L. Eubank		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0207
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University Dallas, Texas 75275		10. PROGRAM ELEMENT. PROJECT. TASK AREA & WORK UNIT NUMBERS NR 042-479
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217		12. REPORT DATE June 1982
		13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report)
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purposes of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Approximation; Density-Quantile Function; Quantile Function; Optimal Grouping; Optimal Spacing; Optimal Stratification

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The relationships between two distributions having the same solutions for problems of optimal spacing selection for the asymptotically best linear un-biased estimator of a location or scale parameter or for problems of optimal stratification for estimation of a population mean are investigated. Easily checked necessary and sufficient conditions under which two distributions have identical solutions to these problems are given in terms of their quantile and density-quantile functions. As an application of these results a quantile domain analog of a theorem due to Adatia and Chan (1981, Scand. Actuar. J.,

(20. Continued)   193-202) on the equivalence of optimal grouping, spacing
and stratification problems is obtained.